



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/752,432	01/07/2004	Jason G. Maxham	53217-014	3709

7590 01/05/2006

MCDERMOTT, WILL & EMERY
600 13th Street, N.W.
Washington, DC 20005-3096

EXAMINER

LIANG, GWEN

ART UNIT PAPER NUMBER

2162

DATE MAILED: 01/05/2006

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Application No. 10/752,432	Applicant(s) MAXHAM ET AL.	
	Examiner GWEN LIANG	Art Unit 2162	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

- 1) ☒ Responsive to communication(s) filed on 13 October 2005.
- 2a) ☐ This action is **FINAL**. 2b) ☒ This action is non-final.
- 3) ☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

- 4) ☒ Claim(s) 1-34 is/are pending in the application.
- 4a) Of the above claim(s) 14-28 is/are withdrawn from consideration.
- 5) ☐ Claim(s) _____ is/are allowed.
- 6) ☒ Claim(s) 1-13 and 29-34 is/are rejected.
- 7) ☐ Claim(s) _____ is/are objected to.
- 8) ☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

- 9) ☐ The specification is objected to by the Examiner.
- 10) ☒ The drawing(s) filed on 07 January 2004 is/are: a) ☐ accepted or b) ☒ objected to by the Examiner.
Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).
Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
- 11) ☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

- 12) ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
- a) ☐ All b) ☐ Some * c) ☐ None of:
1. ☐ Certified copies of the priority documents have been received.
 2. ☐ Certified copies of the priority documents have been received in Application No. _____.
 3. ☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

- | | |
|--|---|
| 1) <input checked="" type="checkbox"/> Notice of References Cited (PTO-892) | 4) <input type="checkbox"/> Interview Summary (PTO-413)
Paper No(s)/Mail Date. _____ |
| 2) <input type="checkbox"/> Notice of Draftsperson's Patent Drawing Review (PTO-948) | 5) <input type="checkbox"/> Notice of Informal Patent Application (PTO-152) |
| 3) <input type="checkbox"/> Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08)
Paper No(s)/Mail Date _____ | 6) <input type="checkbox"/> Other: _____ |

DETAILED ACTION

1. This action is responsive to communications through the applicant's amendment, filed on 10/13/2005.

Drawings

2. This application has been filed with informal drawings (Figures 3-6, 8-10) which are acceptable for examination purposes only. Formal drawings will be required when the application is allowed.

Claim Rejections - 35 USC § 103

3. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negated by the manner in which the invention was made.

4. Claims 1, 4-13, and 29-33 are rejected under 35 U.S.C. 103(a) as being unpatentable over Burrows (U.S. Patent No. 5,745,900), in view of Burkowski ("Retrieval Performance of a Distributed Text Database Utilizing a Parallel Processor Document Server").

5. With respect to claim 1, Burrows discloses a method ...comprising:

- b) creating a fingerprint for each native document (col. 8, lines 16-23, "The FINGERPRINT 255 represents the entire content of the page. The fingerprint 255 can be produced by applying one-way polynomial functions to the digitized content. Typically, the fingerprint is expressed as an integer value. Fingerprinting techniques

Art Unit: 2162

ensure that duplicate pages having identical content have identical fingerprints. With very high probabilities, pages containing different content will have different fingerprints.");

c) de-duplicating each native document in accordance with the fingerprint to produced a de-duplicated plurality of native documents (col. 1, lines 42-45, Therefore, it is desired to provide a technique which minimizes the likelihood that duplicate pages are indexed. The technique should also allow for reindexing as duplicate pages are deleted."; col. 2, lines 41-42, "FIG. 24 shows a process for detecting duplicate pages; FIG. 25 is a flow diagram of a process for deleting pages;"; col. 5, lines 12-14, "The maintenance module 80 also effectively deals with duplicate Web pages containing substantially identical content.");

d) extracting data from each native document; e) associating extracted data with a corresponding native document (col. 5, lines 33-38, "A page 200 can be defined as a data record including a collection of portions of information or "words" having a common database address, e.g., a URL. This means that a page can effectively be a data record of any size, from a single word, to many words, e.g., a large document, a data file, a book, a program, or a sequence of images."; col. 11, line 66 – col. 12, line 7, "The samples are used to generate summary entries 925 in the second level summary data structure 72. Each summary entry 925 includes the word 926 associated with the sample, and the sampled location associated with the word. In addition, the summary entry 925 includes a pointer 928 of the next entry in the compressed data structure 71 following the sampled entry. The summary data structure 72 can also be mapped into

fixed size blocks or disk files to fully populate the summary data structure 72.", wherein the words are extracted from the native documents); and

However Burrows does not explicitly disclose the step of distributing the de-duplicated plurality of native documents and extracted data substantially equally amongst a plurality of nodes of the document management computer system.

Burkowski teaches a method comprising distributing documents and extracted data substantially equally amongst a plurality of nodes of the document management computer system (page 73, section "Uniform Distribution, "The database contents are distributed in a very straightforward fashion; document j and the contribution of this document to the index facility are both loaded into disk $j(\text{mod } N)$. Thus, each disk will contain $1/N$ of all the documents in the database and will also support the index (inverted lists [1], surrogate file [4], [5], [7] or subdivided surrogate files [3]) associated with that particular subset of the full document collection. In this case the server portion of the overall search functionality is divided across N basically independent subsystems").

It would have been obvious to one having ordinary skill in the art at the time the invention was made to adopt a method of distributing documents substantially equally amongst a plurality of nodes as disclosed by Burkowski to load the de-duplicated native documents and extracted data as disclosed in Burrows.. The strategy for uniform distribution of server functionality provides a very reasonable approach when establishing a text retrieval system on a set of parallel servers. Performance can approach linear speed up depending on the overheads involved in handling the index

(page 78, section "Conclusion"). One of ordinary skill in the art would be motivated to make the aforementioned combination with reasonable expectation of success.

Claim 4 is rejected for the reasons set forth hereinabove for claim 1 and furthermore Burrows discloses a method wherein step (c) further comprises comparing the fingerprint of each native document with a plurality of fingerprints comprised of the fingerprints for each native document to be uploaded (col. 28, lines 40-47).

Claim 5 is rejected for the reasons set forth hereinabove for claim 1 and furthermore Burrows discloses a method wherein step (c) further comprises comparing the fingerprint of each native document with at least one fingerprint corresponding to a native document stored in the document management computer system (col. 28, lines 40-47).

Claim 6 is rejected for the reasons set forth hereinabove for claim 4 and furthermore Burrows discloses a method comprising discarding native documents that are determined to be the same in accordance with the comparison of fingerprints (Title; col. 1, lines 42-45; col. 8, lines 16-23).

Claim 7 is rejected for the reasons set forth hereinabove for claim 5 and furthermore Burrows discloses a method comprising discarding native documents that are determined to be the same in accordance with the comparison of fingerprints (Title; col. 1, lines 42-45; col. 8, lines 16-23).

Claim 8 is rejected for the reasons set forth hereinabove for claim 1 and furthermore Burrows discloses a method wherein step (d) further comprises creating at least one data file corresponding to the extracted data for each native document (col. 11, line 66 – col. 12, line 7).

Claim 9 is rejected for the reasons set forth hereinabove for claim 1 and furthermore Burrows discloses a method wherein step (d) further comprises creating a plurality of data files corresponding to the extracted data for each native document (col. 11, line 66 – col. 12, line 7).

Claim 10 is rejected for the reasons set forth hereinabove for claim 9 and furthermore Burrows discloses a method wherein the plurality of data files includes files selected from a group consisting of a text file, a meta data file, an XML file and a HTML file (col. 8, line 66 – col. 9, line 8).

Claim 11 is rejected for the reasons set forth hereinabove for claim 10 and furthermore Burrows discloses a method wherein in step (e), a data table is created for at least one native document for defining an association with the plurality of data files (col. 14, lines 35-40).

Claim 12 is rejected for the reasons set forth hereinabove for claim 1 and furthermore Burrows discloses a method wherein in step (e), a data table is created for at least one native document for defining an association with extracted data (col. 14, lines 35-40).

Claim 13 is rejected on grounds corresponding to the reasons given above for claim 1.

Claim 29 is rejected for the reasons set forth hereinabove for claim 1 and furthermore the combination of Burrows and Burkowski discloses a system comprising a computer in communication with the plurality of computer nodes for receiving a plurality of input files to be uploaded to the plurality of computer nodes (Burrows, col. 2, lines 51-56, "FIG. 1 shows a distributed computer system 100 including a database to be indexed. The distributed system 100 includes client computers 110 connected to server computers (sites) 120 via a network 130. The network 130 can use Internet communications protocols (IP) to allow the clients 110 to communicate with the servers 120."; Burkowski, page 73, section "Uniform Distribution, "The database contents are distributed in a very straightforward fashion; document j and the contribution of this document to the index facility are both loaded into disk $j(\text{mod } N)$. Thus, each disk will contain $1/N$ of all the documents in the database").

The subject matter of claims 30 and 33 are rejected in the analysis above in claim 1, and therefore these claims are rejected on that basis.

The subject matter of claims 31 and 32 are rejected in the analysis above in claims 8 and 10 respectively, and therefore these claims are rejected on that basis.

Art Unit: 2162

6. Claim 2 is rejected under 35 U.S.C. 103(a) as being unpatentable over Burrows (U.S. Patent No. 5,745,900), in view of Burkowski ("Retrieval Performance of a Distributed Text Database Utilizing a Parallel Processor Document Server"), and further in view of Okabe et al., "Okabe " (U.S. Publication No. 2001/0025287)

Claim 2 is rejected for the reasons set forth hereinabove for claim 1. However the combination of Burrows and Burkowski does not explicitly teach a method comprising the step of extracting native document(s) included in the plurality of documents from an archive file.

Okabe teaches the step of extracting native document(s) included in the plurality of documents from an archive file (page 6, section [0077]).

It would have been obvious to one having ordinary skill in the art at the time the invention was made to incorporate a step of extracting native document(s) included in the plurality of documents from an archive file as disclosed by Okabe into the method of managing a plurality of native documents as disclosed in the combination of Burrows and Burkowski. The motivation obviously is to obtain documents from the archive through the extraction (page 6, section [0077]). One of ordinary skill in the art would be motivated to make the aforementioned combination with reasonable expectation of success.

7. Claim 3 is rejected under 35 U.S.C. 103(a) as being unpatentable over Burrows (U.S. Patent No. 5,745,900), in view of Burkowski ("Retrieval Performance of a

Distributed Text Database Utilizing a Parallel Processor Document Server”), and further in view of Zabetian (U.S. Publication No. 2001/0011350).

Claim 3 is rejected for the reasons set forth hereinabove for claim 1. However the combination of Burrows and Burkowski does not explicitly teach a method wherein the fingerprint for each native document is created using a MD5 checksum.

Zabetian teaches a method wherein the fingerprint for each native document is created using a MD5 checksum (page 4, section [0037]).

It would have been obvious to one having ordinary skill in the art at the time the invention was made to incorporate a method wherein the fingerprint for each native document is created using a MD5 checksum as disclosed by Zabetian into the method of creating a fingerprint for each native document as disclosed in the combination of Burrows and Burkowski, where a tamper proof checksum algorithm is desired, MD5 with DES encryption can be used (MD5-DES) (page 4, section [0037]). One of ordinary skill in the art would be motivated to make the aforementioned combination with reasonable expectation of success.

8. Claim 34 is rejected under 35 U.S.C. 103(a) as being unpatentable over Burrows (U.S. Patent No. 5,745,900), in view of Burkowski (“Retrieval Performance of a Distributed Text Database Utilizing a Parallel Processor Document Server”), and further in view of Froessl (U.S. Patent No.5,444,840).

With respect to claim 34, Burrows discloses a system ...comprising:

a PC type computer a PC type computer connected in a parallel cluster (col. 2, lines 51-56; col. 15, lines 6-14),

said computer using an operating system that generates a fingerprint for each document (col. 8, lines 16-23, "The FINGERPRINT 255 represents the entire content of the page. The fingerprint 255 can be produced by applying one-way polynomial functions to the digitized content. Typically, the fingerprint is expressed as an integer value. Fingerprinting techniques ensure that duplicate pages having identical content have identical fingerprints. With very high probabilities, pages containing different content will have different fingerprints.");

where each document is identified by its file extension (col. 7, lines 58-65, "For example, the page 200 of FIG. 4 can have associated page attributes 250. Page attributes 250 can include □ADDRESS□ 251, □DESCRIPTION□ 252, □SIZE□ 253, □DATE□ 254, □FINGERPRINT□ 255, □TYPE□ 256, and □END_PAGE□ 257, for example. The symbol "□," represents one or more characters which cannot be confused with the characters normally found in words, for example "space," "underscore," and "space" (sp_sp)"; col. 8, lines 24-25, "The TYPE attribute 256 may distinguish pages having different multimedia content or formatting characteristics"; Figure 4, element 256; It is well known to an ordinary skill in the art that each file has a file extension, which indicate the type of the file or document);

and given a unique identification number (col. 26, lines 4-6, "Each entry 2201 includes an identification (page_id) 2210 of a qualified page"),

each of a plurality of documents having at least one of either meta-data, text or attachments that are indexed for web-based retrieval from the cluster (col. 8, line 66 – col. 9, line 8, "Attribute values or metawords can be generated for portions of a page. For example, the words of the field 230 may be the "title" of the page 200. In this case the "title" has a first word 231 and a last word 239. In "html" pages, the titles can be expressly noted. In other types of text, the title may be deduced from the relative placement of the words on the page, for example, first line centered. For titles, the parsing module 30 can generate a "BEGIN_TITLE" pair and an "END_TITLE" pair to be respectively associated with the locations of the first and last words of the title."; col. 3, line 31, "means for indexing the parsed pages"; col. 5, lines 26-27, "In the index 70 each word is stored as a "literal" or character based value"; col. 8, lines 44-46, "By inserting the . "END_PAGE" attribute value in the index 70 as a metaword, searching the index as described below can be more efficient"; col. 9, lines 43-46, "the indexing module 40 generates an index 70 of the content of the records or pages 200. The internal data structures 71-73 of the index 70 are now described first with reference to FIG. 6");

said plurality of documents are de-duplicated in accordance with its fingerprint (col. 1, lines 42-45, "Therefore, it is desired to provide a technique which minimizes the likelihood that duplicate pages are indexed. The technique should also allow for reindexing as duplicate pages are deleted."; col. 2, lines 41-42, "FIG. 24 shows a process for detecting duplicate pages; FIG. 25 is a flow diagram of a process for

Art Unit: 2162

deleting pages;"; col. 5, lines 12-14, "The maintenance module 80 also effectively deals with duplicate Web pages containing substantially identical content.");

said plurality of documents forming a cluster data base that is web-searchable by use of a predetermined descriptive term (col. 3, lines 28-33, "In order to identify pages of interest among the millions of pages which are available on the Web, a search engine 140 is provided. The search engine 140 includes means for parsing the pages, means for indexing the parsed pages, means for searching the index, and means for presenting information about the pages 200 located.").

However Burrows does not explicitly teach a system that stores electronic documents substantially equally in number throughout the cluster and a system wherein each document is converted to ASCII text.

Burkowski teaches a system that stores electronic documents substantially equally in number throughout the cluster (page 73, section "Uniform Distribution, "The database contents are distributed in a very straightforward fashion; document j and the contribution of this document to the index facility are both loaded into disk $j(\text{mod } N)$). Thus, each disk will contain $1/N$ of all the documents in the database and will also support the index (inverted lists [1], surrogate file [4], [5], [7] or subdivided surrogate files [3]) associated with that particular subset of the full document collection. In this case the server portion of the overall search functionality is divided across N basically independent subsystems").

It would have been obvious to one having ordinary skill in the art at the time the invention was made to adopt a system that stores electronic documents substantially

Art Unit: 2162

equally in number throughout the cluster as disclosed by Burkowski to store the de-duplicated as disclosed in Burrows. The strategy for uniform distribution of server functionality provides a very reasonable approach when establishing a text retrieval system on a set of parallel servers. Performance can approach linear speed up depending on the overheads involved in handling the index (page 78, section "Conclusion"). One of ordinary skill in the art would be motivated to make the aforementioned combination with reasonable expectation of success.

However the combination of Burrows and Burkowski does not explicitly teach a system wherein each document is converted to ASCII text.

Froessl teaches a system wherein each document is converted to ASCII text (Abstract, "In one embodiment, the image representation is converted into code (ASCII)").

It would have been obvious to one having ordinary skill in the art at the time the invention was made to incorporate a system where each document is converted to ASCII text as disclosed by Froessl into the an electronic document management system as disclosed in the combination of Burrows and Burkowski. Systems of this type allow full-text code searches to be conducted for words which appear in the documents. An advantage of this type of system is that indexing is not absolutely required because the full text of each document can be searched, allowing a document dealing with a specific topic or naming a specific person to be located without having to be concerned with whether the topic or person was named in the index (col. 1 lines 41-49). One of

Art Unit: 2162

ordinary skill in the art would be motivated to make the aforementioned combination with reasonable expectation of success.

Response to Arguments

9. Applicant's arguments with respect to all the pending claims have been considered but are moot in view of the new ground(s) of rejection.


Contact Information

Any inquiry concerning this communication or earlier communications from the examiner should be directed to GWEN LIANG whose telephone number is 571-272-4038. The examiner can normally be reached on 9:30 A.M. - 5:30 P.M. Monday and Thursday.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, JOHN BREENE can be reached on 571-272-4107. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

29 December 2005
G.L.


Primary Examiner
Art Unit 2167